



Práctica RIWS – Crawler de componentes

Antón Gendra Rodríguez
Adrián Delfín Mella Castelo
Jacobo Fiaño Rodríguez
Juan Andrés Justo Armesto
18-11-2022

Índice

1. Introducción	2
2. Tecnologías usadas.....	3
3. Search casos solucionados	4
4. Solución desarrollada.....	6

1. Introducción

Para la realización de esta práctica decidimos crear una herramienta que permita a los usuarios buscar componentes de ordenador. Esta herramienta debe permitir realizar búsquedas filtradas por determinadas características de los componentes, entre las cuales se encuentra: marca, nombre, tipo de componente, etc.

Hay que tener en cuenta que al ser un proyecto de recuperación de información el objetivo de este no es proveer los componentes en sí como servicio final, sino que obtenemos los datos de diferentes webs que expondremos en la nuestra.

Las webs de las que procede la información son:

- <https://www.coolmod.com/>
- <https://www.pcmontajes.com/>
- <https://www.neobyte.es/>
- <https://www.pcbox.com/>

También se trató de *scrapear* las siguientes páginas, aunque no se obtuvieron resultados satisfactorios, por lo que no se tuvieron en cuenta:

- <https://www.gigabyte.com/>
- <https://www.pccomponentes.com/pccom>
- <https://www.eneba.com/>
- <https://www.vsgamers.es/>
- <https://www.aussar.es/>
- <https://www.fnac.com/>
- <https://www.wipoid.com/>
- <https://www.alternate.es/>
- <https://www.mediamarkt.es/>
- <https://tienda.redcomputer.es/>
-

Para cada componente decidimos obtener las siguientes características:

- Id (name-source)
- Clicks
- Name
- Brand
- Price
- Link
- Source
- Weight
- Height
- Width
- Category
- Storing_capacity

- Power
- Speed
- Latency
- Max_temperature
- Year
- Generation
- Rating
- Socket
- Interface
- Architecture
- Image
- Cores
- Threads

2. Tecnologías usadas

Para el desarrollo del proyecto se usó mayoritariamente Django. Existían dos alternativas para el desarrollo de la interfaz web de manera ágil, no obstante, se usó Django debido a que dos miembros del equipo tenían experiencia previa con el framework, posibilitando un desarrollo mucho más rápido y sencillo.

Se empleó Elasticsearch para el almacenamiento de los datos recopilados de las diferentes fuentes de información. Como primer paso, el *pipeline* crea el índice y lanza los diferentes *Spiders*, acto seguido, cada *spider* recupera la información de su respectiva web. Una vez el *pipeline* recibe los ítems de los *spiders* este los introduce en el Índice realizando una petición POST. Al ejecutar una búsqueda, los controladores de las vistas llaman al método que consigue los datos de Elasticsearch.

Usamos SQLite y JavaScript para el almacenamiento de datos, en concreto se almacena la relación entre los componentes y el número de veces a la que se accede (*clicks*).

Scrapy se emplea para *crawlear* las páginas y obtener los datos de los html que las constituyen.

Se designan links a los cuales se quiere acceder con las reglas allow, se accede a todos los que *macheen* para obtener la lista de componentes de esa página. Una vez obtenemos la lista, por cada elemento, obtenemos su link al que accedemos individualmente (por cuestiones de eficiencia) para obtener su información, haciendo uso del método `response.css()`.

Empleamos Tailwind-css para mejorar el aspecto visual de la web, el motivo principal es que uno de los miembros del equipo tiene experiencia previa con esta tecnología.

3. Search casos solucionados

Para realizar una búsqueda podemos filtrar los componentes por varios parámetros, entre ellos:

- Búsqueda por nombre del artículo.
- Marca, pudiendo seleccionar diversas marcas.
- Rango de precio, valor mínimo y máximo.
- Web de origen (source), pudiendo seleccionar diversas webs.
- Categoría del componente (Placa base, Procesador, memoria RAM, SSD, HDD, tarjeta gráfica, tarjeta de red, tarjeta de sonido, refrigeración, fuente de alimentación, torre).
- Además, se podrá filtrar también características más específicas de cada componente (socket, capacidad máxima, potencia, máx. temperatura, velocidad, peso, altura, ancho).

Para realizar la búsqueda en Elasticsearch utilizamos una boolean query, formada por 2 queries, en primer lugar, una query de tipo *filter* mediante la cual se hace un filtrado de los componentes según la categoría introducida por el usuario. Junto a esta, se añade un query booleana de tipo *must* formada por otras 3 queries, una de tipo *should* en la que se introduce dos queries para filtrar por la web de origen y la marca. La segunda query es un *filter* mediante la cual se filtra por el rango de precio y, por último, una tercera query de tipo *must* en la que se introducen los filtros de las características concretas para el componente buscado.

```
query = {
  "bool": {
    "filter": [
      { "term": { "category": "processor" } }
    ],
    "must": [
      { "bool": {
        "should": [
          { "term": { "brand": "Kingstone" } },
          { "term": { "brand": "Gigabyte" } }
        ]
      } },
      { "range": {
        "price": { "gte": 15, "lte": 150 }
      } },
      {
        "nested": {
          "path": "characteristics",
          "query": {
            "bool": {
              "must": [
                { "match": { "characteristics.socket": "LG234" } },
                { "match": { "characteristics.storing_capacity": 5000 } }
              ]
            }
          }
        }
      }
    ]
  }
}
```

Ilustración 1. Ejemplo de query de búsqueda

4. Solución desarrollada

Para guardar los componentes en Elasticsearch se hizo uso de un id, el cual está formado por: name-source, que corresponden al nombre del producto y la página web de la que se obtuvo.

A mayores se implementó una funcionalidad la cual cuenta el número de clicks y aumenta la relevancia del componente en los resultados recuperados de cada query.

A continuación, se muestran los resultados obtenidos en cada uno de los *spider* (Ilustración 1, Ilustración 2, Ilustración 3, Ilustración 4):

```
2022-11-18 21:33:42 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.coolmod.com/asus-prime-a320m-e-socket-am4-placa-base/> (referer: https://www.coolmod.com/componentes-pc-placas-base/)
2022-11-18 21:33:42 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_search HTTP/1.1" 200 160
2022-11-18 21:33:42 [elastic_transport.transport] INFO: POST http://localhost:9200/component_index/_search [status:200 duration:0.005s]
2022-11-18 21:33:42 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_doc HTTP/1.1" 201 167
2022-11-18 21:33:42 [elastic_transport.transport] INFO: POST http://localhost:9200/component_index/_doc [status:201 duration:0.014s]
Item añadido, id: Asus Prime A320M-E Socket AM4 - Placa Base
2022-11-18 21:33:42 [scrapy.core.scraper] DEBUG: Scraped from <200 https://www.coolmod.com/asus-prime-a320m-e-socket-am4-placa-base/>
{'brand': 'ASUS',
 'category': 'motherboard',
 'image': 'https://cdn.coolmod.com/images/product/normal/asus-prime-a320m-e-socket-am4-placa-base-001.jpg',
 'link': 'https://www.coolmod.com/asus-prime-a320m-e-socket-am4-placa-base/',
 'name': 'Asus Prime A320M-E Socket AM4 - Placa Base ',
 'price': 97.95,
 'source': 'Coolmod'}
2022-11-18 21:33:47 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.coolmod.com/gigabyte-a320m-s2h-socket-am4-placa-base/> (referer: https://www.coolmod.com/componentes-pc-placas-base/)
2022-11-18 21:33:47 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_search HTTP/1.1" 200 160
2022-11-18 21:33:47 [elastic_transport.transport] INFO: POST http://localhost:9200/component_index/_search [status:200 duration:0.006s]
2022-11-18 21:33:48 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_doc HTTP/1.1" 201 168
2022-11-18 21:33:48 [elastic_transport.transport] INFO: POST http://localhost:9200/component_index/_doc [status:201 duration:0.020s]
Item añadido, id: Gigabyte A320M-S2H Socket AM4 - Placa Base
2022-11-18 21:33:48 [scrapy.core.scraper] DEBUG: Scraped from <200 https://www.coolmod.com/gigabyte-a320m-s2h-socket-am4-placa-base/>
{'brand': 'GIGABYTE',
 'category': 'motherboard',
 'image': 'https://cdn.coolmod.com/images/product/normal/gigabyte-a320m-s2h-socket-am4-placa-base-001.jpg',
 'link': 'https://www.coolmod.com/gigabyte-a320m-s2h-socket-am4-placa-base/',
 'name': 'Gigabyte A320M-S2H Socket AM4 - Placa Base ',
 'price': 49.95,
 'socket': '\n\t',
 'source': 'Coolmod'}
```

Ilustración 2. Spider CoolMod

```
2022-11-18 21:39:50 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.neobyte.es/asus-prime-x670e-pro-wi-fi-placa-base-amd-am5-15310.html> (referer: https://www.neobyte.es/placas-base-106)
2022-11-18 21:39:50 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_search HTTP/1.1" 200 160
2022-11-18 21:39:50 [elastic_transport.transport] INFO: POST http://localhost:9200/component_index/_search [status:200 duration:0.006s]
2022-11-18 21:39:50 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_doc HTTP/1.1" 201 168
2022-11-18 21:39:50 [elastic_transport.transport] INFO: POST http://localhost:9200/component_index/_doc [status:201 duration:0.013s]
Item añadido, id: Asus Prime X670E-Pro Wi-Fi - Placa base AMD AM5
2022-11-18 21:39:50 [scrapy.core.scraper] DEBUG: Scraped from <200 https://www.neobyte.es/asus-prime-x670e-pro-wi-fi-placa-base-amd-am5-15310.html>
{'brand': 'AMD',
 'category': 'motherboard',
 'image': 'https://www.neobyte.es/79987-home_default/asus-prime-x670e-pro-wi-fi-placa-base-amd-am5.jpg',
 'link': 'https://www.neobyte.es/asus-prime-x670e-pro-wi-fi-placa-base-amd-am5-15310.html',
 'name': 'Asus Prime X670E-Pro Wi-Fi - Placa base AMD AM5',
 'price': 375.9,
 'rating': None,
 'socket': 'amd socket am5 for amd ryzen™ 7000 series desktop processors',
 'source': 'Neobyte'}
2022-11-18 21:39:56 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.neobyte.es/placa-base-rog-strix-x299-e-gaming-ii-5811.html> (referer: https://www.neobyte.es/placas-base-106)
2022-11-18 21:39:56 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_search HTTP/1.1" 200 160
2022-11-18 21:39:56 [elastic_transport.transport] INFO: POST http://localhost:9200/component_index/_search [status:200 duration:0.006s]
2022-11-18 21:39:56 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_doc HTTP/1.1" 201 168
2022-11-18 21:39:56 [elastic_transport.transport] INFO: POST http://localhost:9200/component_index/_doc [status:201 duration:0.014s]
Item añadido, id: Asus ROG Strix X299-E Gaming II - Placa base Intel 2066
2022-11-18 21:39:56 [scrapy.core.scraper] DEBUG: Scraped from <200 https://www.neobyte.es/placa-base-rog-strix-x299-e-gaming-ii-5811.html>
{'brand': 'INTEL',
 'category': 'motherboard',
 'image': 'https://www.neobyte.es/31557-home_default/placa-base-rog-strix-x299-e-gaming-ii.jpg',
 'link': 'https://www.neobyte.es/placa-base-rog-strix-x299-e-gaming-ii-5811.html',
 'name': 'Asus ROG Strix X299-E Gaming II - Placa base Intel 2066',
 'price': 489.94,
 'rating': None,
 'socket': 'intel® socket 2066 core™ x-series processors',
 'source': 'Neobyte'}
```

Ilustración 3. Spider NeoByte

```

2022-11-18 21:41:44 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.pcbbox.com//sa400s37-240g-kingston--a400--ssd-240gb-2-5---500mb-s-6gbt-s--serial-ata-iii/p> (referer: https://www.pcbbox.com/component-es-de-ordenador)
2022-11-18 21:41:44 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_search HTTP/1.1" 200 160
2022-11-18 21:41:44 [elastic.transport.transport] INFO: POST http://localhost:9200/component_index/_search [status:200 duration:0.005s]
2022-11-18 21:41:44 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_doc HTTP/1.1" 201 168
2022-11-18 21:41:44 [elastic.transport.transport] INFO: POST http://localhost:9200/component_index/_doc [status:201 duration:0.015s]
Item añadido, id: DISCO DURO 240GB 2.5" KINGSTON SSD SATA3 A400
2022-11-18 21:41:44 [scrapy.core.scrapy] DEBUG: Scraped from <200 https://www.pcbbox.com//sa400s37-240g-kingston--a400--ssd-240gb-2-5---500mb-s-6gbt-s--serial-ata-iii/p>
{'brand': 'KINGSTON',
 'category': 'SSD',
 'height': 7.0,
 'image': 'https://pbox.vtexassets.com/arquivos/ids/396366-300-300?v=1755191136&width=300&height=300&aspect=true',
 'interface': 'Serial ATA III',
 'link': 'https://www.pcbbox.com//sa400s37-240g-kingston--a400--ssd-240gb-2-5---500mb-s-6gbt-s--serial-ata-iii/p',
 'name': 'DISCO DURO 240GB 2.5" KINGSTON SSD SATA3 A400',
 'price': 23.75,
 'source': 'PcBox',
 'speed': 500.0,
 'storing_capacity': 240.0,
 'type': 'TLC',
 'weight': 41.0,
 'width': 100.0}
2022-11-18 21:41:49 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.pcbbox.com//nshumera600hubz-fuente-alimentacion-600w-nox-hummer-alpha-80--bronzep> (referer: https://www.pcbbox.com/componentes-de-ordenador)
2022-11-18 21:41:49 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_search HTTP/1.1" 200 160
2022-11-18 21:41:49 [elastic.transport.transport] INFO: POST http://localhost:9200/component_index/_search [status:200 duration:0.009s]
2022-11-18 21:41:49 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_doc HTTP/1.1" 201 168
2022-11-18 21:41:49 [elastic.transport.transport] INFO: POST http://localhost:9200/component_index/_doc [status:201 duration:0.018s]
Item añadido, id: FUENTE ALIMENTACION 600W NOX HUMMER ALPHA 80w BRONZE
2022-11-18 21:41:49 [scrapy.core.scrapy] DEBUG: Scraped from <200 https://www.pcbbox.com//nshumera600hubz-fuente-alimentacion-600w-nox-hummer-alpha-80--bronzep>
{'brand': 'NOX HUMMER',
 'category': 'power-source',
 'height': 86.0,
 'image': 'https://pbox.vtexassets.com/arquivos/ids/486937-300-300?v=1755172788&width=300&height=300&aspect=true',
 'link': 'https://www.pcbbox.com//nshumera600hubz-fuente-alimentacion-600w-nox-hummer-alpha-80--bronzep',
 'name': 'FUENTE ALIMENTACION 600W NOX HUMMER ALPHA 80w BRONZE',
 'power': 600.0,
 'price': 54.9,
 'source': 'PcBox',
 'weight': 2.01,
 'width': 150.0}

```

Ilustración 4. Spider PcBox

```

2022-11-18 22:16:39 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.pcmontajes.com/componentes-cajas-pc/28126-antec-nx200m-cristal-templado-caja-torre-0761345810272.html> (referer: https://www.pcmontajes.com/47-componentes-cajas-pc)
2022-11-18 22:16:39 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_search HTTP/1.1" 200 160
2022-11-18 22:16:39 [elastic.transport.transport] INFO: POST http://localhost:9200/component_index/_search [status:200 duration:0.010s]
2022-11-18 22:16:39 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_doc HTTP/1.1" 201 168
2022-11-18 22:16:39 [elastic.transport.transport] INFO: POST http://localhost:9200/component_index/_doc [status:201 duration:0.026s]
Item añadido, id: Antec NX200M Cristal Templado - Caja Torre
2022-11-18 22:16:39 [scrapy.core.scrapy] DEBUG: Scraped from <200 https://www.pcmontajes.com/componentes-cajas-pc/28126-antec-nx200m-cristal-templado-caja-torre-0761345810272.html>
{'brand': 'ANTEC',
 'category': 'tower',
 'height': 390.0,
 'image': 'https://www.pcmontajes.com/94289-home_default/antec-nx200m-cristal-templado-caja-torre.jpg',
 'link': 'https://www.pcmontajes.com/componentes-cajas-pc/28126-antec-nx200m-cristal-templado-caja-torre-0761345810272.html',
 'name': 'Antec NX200M Cristal Templado - Caja Torre',
 'price': '48.61',
 'source': 'pcmontajes',
 'weight': 3.51,
 'width': 200.0}
2022-11-18 22:16:41 [scrapy.extensions.logstats] INFO: Crawled 24 pages (at 11 pages/min), scraped 11 items (at 11 items/min)
2022-11-18 22:16:45 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.pcmontajes.com/componentes-placas-base/26260-placa-base-asus-am4-tuf-gaming-b550m-e-4711081173397.html> (referer: https://www.pcmontajes.com/36-componentes-placas-base)
2022-11-18 22:16:45 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_search HTTP/1.1" 200 160
2022-11-18 22:16:45 [elastic.transport.transport] INFO: POST http://localhost:9200/component_index/_search [status:200 duration:0.011s]
2022-11-18 22:16:46 [urllib3.connectionpool] DEBUG: http://localhost:9200 "POST /component_index/_doc HTTP/1.1" 201 168
2022-11-18 22:16:46 [elastic.transport.transport] INFO: POST http://localhost:9200/component_index/_doc [status:201 duration:0.031s]
Item añadido, id: PLACA BASE ASUS AM4 TUF GAMING B550M-E
2022-11-18 22:16:46 [scrapy.core.scrapy] DEBUG: Scraped from <200 https://www.pcmontajes.com/componentes-placas-base/26260-placa-base-asus-am4-tuf-gaming-b550m-e-4711081173397.html>
{'brand': 'ASUS',
 'category': 'motherboard',
 'image': 'https://www.pcmontajes.com/89189-home_default/placa-base-asus-am4-tuf-gaming-b550m-e.jpg',
 'link': 'https://www.pcmontajes.com/componentes-placas-base/26260-placa-base-asus-am4-tuf-gaming-b550m-e-4711081173397.html',
 'name': 'PLACA BASE ASUS AM4 TUF GAMING B550M-E',
 'price': '148.31',
 'socket': 'Zócalo AM4',
 'source': 'pcmontajes',
 'storing_capacity': 128.0,
 'width': 244.0}

```

Ilustración 5. Spider PcMontajes

En las Ilustraciones 5 y 6, se muestra la pantalla principal de la interfaz web, en ella podremos realizar búsquedas filtrando por nombre o seleccionar una de las cards que representan cada una de las categorías.

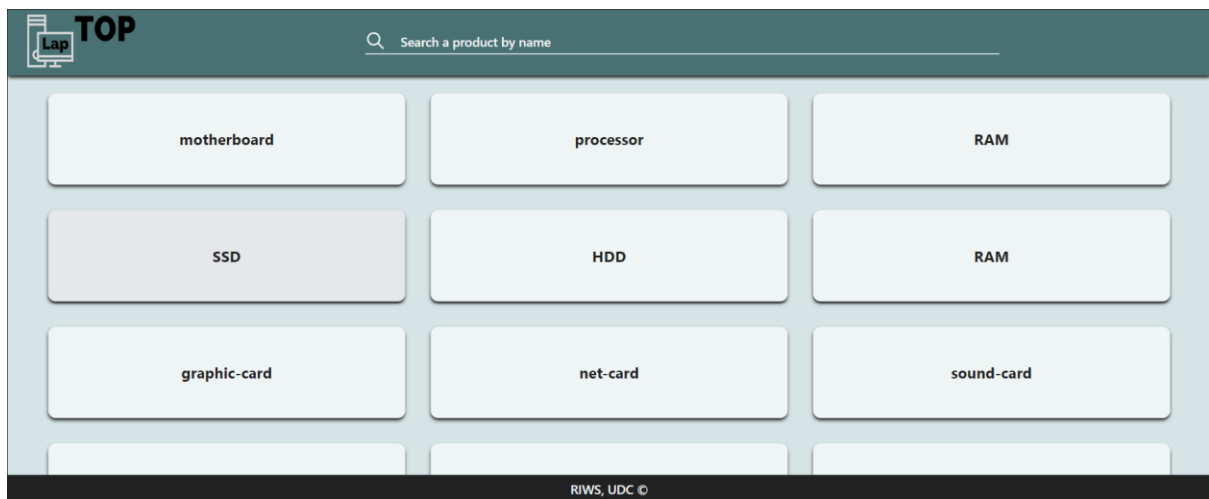


Ilustración 6. Pantalla principal

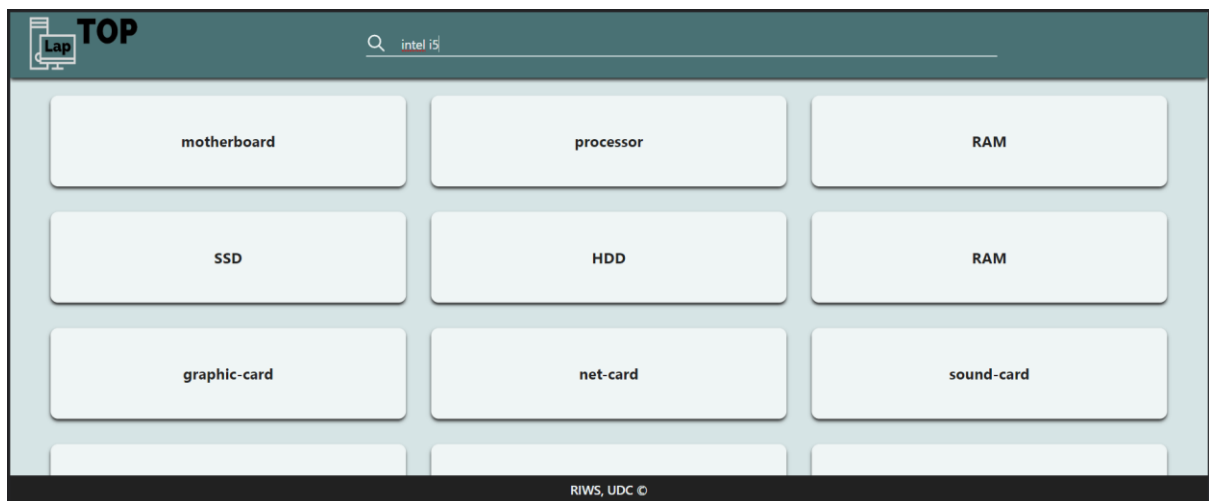


Ilustración 7. Ejemplo de búsqueda





En las Ilustraciones 7, 8 y 9 se muestra la lista de los componentes que coincidieron con la búsqueda. Al *clickar* en una de las tarjetas que representan cada uno de los elementos se redireccionará a la web de este. Además, en esta pantalla podremos aplicar varios filtros que encontramos en la parte superior de la Ilustración 8.

Lap TOP

intel Min. Price Max. Price All categories

All brands: AMD, INTEL, ASUS
All sources: Coolmod, Neobyte, PcBox

Apply filters More filters

 <p>Intel Core i5-10400F ... 135.9 € INTEL Source: Neobyte Clicks: 0</p>	 <p>Intel Core i5-10600 4... 280.95 € INTEL Source: Coolmod Clicks: 0</p>	 <p>Intel Core i5-9500 4... 151.95 € INTEL Source: Coolmod Clicks: 0</p>	 <p>PLACA INTEL CORE i... 84.83 € INTEL Source: PcBox Clicks: 0</p>
---	--	---	--

RIWS, UDC ©

Ilustración 8. Resultado de búsqueda

Lap TOP

intel 205.9 420 processor





LGA 1200 Storing Capacity Power Max. Temperature

Speed Weight Height Width

AMD, INTEL, ASUS, GIGABYTE, MSI, NZXT

All sources: Coolmod, Neobyte, PcBox, PcMontajes

Apply filters

 <p>Intel Core i5-10400F ... 135.9 € INTEL Source: Neobyte Clicks: 0</p>	 <p>Intel Core i5-10600 4... 280.95 € INTEL Source: Coolmod Clicks: 0</p>	 <p>Intel Core i5-9500 4... 151.95 € INTEL Source: Coolmod Clicks: 0</p>	 <p>PLACA INTEL CORE i... 84.83 € INTEL Source: PcBox Clicks: 0</p>
--	---	--	---

RIWS, UDC ©

Ilustración 9. Búsqueda con filtros avanzados

Lap TOP

intel 205.9 420 processor





LGA 1200 Storing Capacity Power Max. Temperature

Speed Weight Height Width

All brands: AMD, INTEL, ASUS, GIGABYTE, MSI

All sources: Coolmod, Neobyte, PcBox, PcMontajes

Apply filters

 <p>Intel Core i7-10700F ... 298.89 € INTEL Source: Neobyte Clicks: 0</p>	 <p>Intel Core i7-12700K ... 399.89 € INTEL Source: Neobyte Clicks: 0</p>	 <p>Intel Core i7-10700K ... 359.95 € INTEL Source: Coolmod Clicks: 0</p>	 <p>Intel Core i5-10600 4... 280.95 € INTEL Source: Coolmod Clicks: 0</p>
--	--	--	--

RIWS, UDC ©

Ilustración 10. Resultado de búsqueda con filtros avanzados

Como podemos ver en la Ilustración 10, los componentes se ordenan según el número de clicks, este número se puede ver en la parte inferior de cada una de las tarjetas.

The screenshot shows a web interface for searching laptop components. At the top, there's a navigation bar with a 'Lap TOP' logo, filters for 'intel', '205,9', '420', and 'processor', and lists of 'All brands' (AMD, INTEL, ASUS) and 'All sources' (Coolmod, Neobyte, PcBox). There are 'Apply filters' and 'More filters' buttons. Below the navigation bar, four product cards are displayed, each representing an Intel processor. Each card includes a product image, the processor name, the price in Euros, the source, and the number of clicks. The cards are sorted by the number of clicks in descending order.

Processor	Price (€)	Source	Clicks
Intel Core i7-10700F ...	298.89 €	Neobyte	8
Intel Core i7-12700K ...	399.89 €	Neobyte	5
Intel Core i5-10600 4...	280.95 €	Coolmod	1
Intel Core i7-10700K ...	359.95 €	Coolmod	0

RIWS, UDC ©

Ilustración 11. Componentes ordenados por número de clicks